Homework 7:

About

This homework focuses on implementing a pipeline predicting the rate of reviews.

Due

Monday 3/27/19, 11:59 PM CST

Goal

External Libraries

You may find the following packages/functions/commands are helpful:

- `pandas`
- `stopwords` in `nltk.corpus`
- `string.punctuation`
- `sklearn`

Problem

At http://courses.engr.illinois.edu/cs498aml/sp2019/homeworks/yelp_2k.csv (http://courses.engr.illinois.edu/cs498aml/sp2019/homeworks/yelp_2k.csv) you will find a dataset of Yelp reviews. The original dataset (https://www.kaggle.com/yelp-dataset/yelp-dataset/version/4) contains 5,261,668 reviews and we select 2000 from them, where half of them for reviews with 1 and 5 stars respectively.

1. Download and import the dataset. And then extract `text` and `stars` columns as your `X` (data) and `y` (label). You may find `pandas` package would be helpful.

2. First glance
   For each class, plot the histogram distribution of the length of the review.

3. Preprocessing and tokenization.
   For each review, you need to
   1). Remove all the punctuations.

2). Convert to lower case.
3). Remove all stopwords.
4). Convert to a list of words.

4. Vectorization / Bag of Words (BoW)
   you should write your own code for this part.
   For example, if you have two sentences:

- He is a graduate student.
- She is a faculty here.

Then the final BoW representation will be in the following. (note that the order of the word does not matter, you could choose any order you like):

He	She	is	a	graduate	faculty	here	student
1	0	1	1	1	0	0	1
0	1	1	1	0	1	1	0

5. Classification
   Now you have 2,000 BoW data and labels. In this part, you need to
   1). Randomly split 80% of the dataset to train and the rest to test.
   2). Build **two** different classifiers (one of them must be logistic regression). Train them on **the same** training dataset. You can use whichever classification method and packages here.
   3). Test two classifiers on **the same** testing dataset and print out the corresponding confusion matrix and accuracy.

6. Analysis
   Compare two models. Which model do you think is better? Explain why you think this model works better.
   a. Plot distributions of score grouped by label
   b. Plot Precision-Recall curve (for AMO students: please watch AML video with relevant discussion)

Submission

Submission will be through gradescope (https://www.gradescope.com/). Your submission should be a PDF with the following pages.

1. (**5 points**) Page 1: First-glance analysis. Plot the distribution of comment length.

2. (**20 points**) Page 2: tokenization preprocessing. Put the snippet of your code here and give an example result of one comment.

3. (**20 points**) Page 3: vectorization preprocessing. Put the snippet of your code here and give an example result of one comment.

4. (**20 points**) Page 4: classifier one. Train classifier on the training dataset and report the accuracy on the test dataset.

5. (**20 points**) Page 5: classifier two. Use **the same** train-test split in the previous step. Train this classifier on the training dataset and report the accuracy on the test dataset.

6. (**10 points**) Page 6: result analysis. Which classifier is better? Why?

7. (**5 points**) Page 7: code.